Vision Matters When It Should: Sanity Checking Multimodal Machine Translation Models

Jiaoda Li[1] **Duygu Ataman**[2*] **Rico Sennrich**[3,4]

[1]ETH Zürich
[2]New York University
[3]University of Zürich
[4]University of Edinburgh
jiaoda.li@inf.ethz.ch ataman@nyu.edu sennrich@cl.uzh.ch

Abstract

Multimodal machine translation (MMT) systems have been shown to outperform their text-only neural machine translation (NMT) counterparts when visual context is available. However, recent studies have also shown that the performance of MMT models is only marginally impacted when the associated image is replaced with an unrelated image or noise, which suggests that the visual context might not be exploited by the model at all. We hypothesize that this might be caused by the nature of the commonly used evaluation benchmark, also known as Multi30K, where the translations of image captions were prepared without actually showing the images to human translators. In this paper, we present a qualitative study that examines the role of datasets in stimulating the leverage of visual modality and we propose methods to highlight the importance of visual signals in the datasets which demonstrate improvements in reliance of models on the source images. Our findings suggest the research on effective MMT architectures is currently impaired by the lack of suitable datasets and careful consideration must be taken in creation of future MMT datasets, for which we also provide useful insights.[1]

1 Introduction

Multimodal machine translation (MMT) aims to improve machine translation by resolving certain contextual ambiguities with the aid of other modalities such as vision, and have shown promising integration in conventional neural machine translation (NMT) models (Specia et al., 2016). On the other hand, recent studies reported some conflicting results regarding how the additional visual information is exploited by the models for generating higher-quality translations. A number of MMT models (Calixto et al., 2017; Helcl et al., 2018; Ive et al., 2019; Lin et al., 2020; Yin et al., 2020) have been proposed which showed improvements over text-only models, whereas Lala et al. (2018); Barrault et al. (2018); Raunak et al. (2019) observed that the multimodal integration did not make a big difference quantitatively or qualitatively. Following experimental work showed that replacing the images in image-caption pairs with incongruent images (Elliott, 2018) or even random noise (Wu et al., 2021) might still result in similar performance of multimodal models. In light of these results, Wu et al. (2021) suggested that gains in quality might merely be due to a regularization effect and the images may not actually be exploited by models during the translation task.

In this paper, we investigate the role of the evaluation benchmark in model performance and whether its tendency to ignore visual information in the input could be a consequence of the nature of the dataset. The most widely-used dataset for MMT is Multi30K (Elliott et al., 2016, 2017; Barrault et al., 2018), which extends the Flickr30K dataset (Young et al., 2014) to German, French, and Czech translations. Captions were translated without access to images, and it is posited that this heavily biases MMT models towards only relying on textual input (Elliott, 2018). MMT models may well be capable of using visual signals, but will only learn to do so if the visual context provides information beyond the text. For instance, the English word "wall" can be translated into German as either "Wand" (wall inside of a building) or "Mauer" (wall outside of a building), but we find that reference translations in Multi30k are not always congruent with images.

A number of efforts have been put into creating datasets where correct translations are only possible in the presence of images. Caglayan et al. (2019) degrade the Multi30K dataset to hide away crucial information in the source sentence, includ-

[*] Work done while at the University of Zürich.

[1]Our code and data are available at: https://github.com/jiaodali/vision-matters-when-it-should.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8556–8562
November 7–11, 2021.

ing color, head nouns, and suffixes. Similarly, Wu et al. (2021) mask high-frequency words in Multi30K. Multisense (Gella et al., 2019) collects sentences whose verbs have cross-lingual sense ambiguities. However, due to the high cost of data collection, datasets of such kind are often limited in size. MultiSubs (Wang et al., 2021) is another related dataset. which is primarily used for lexical translation because the images are retrieved to align with text fragments rather than whole sentences.

In this work, we propose two methods to necessitate the visual context — back-translation from a gender-neutral language (e.g. Turkish) and word dropout in the source sentence. They are simple and cheap to implement, allowing them to be applied on much larger datasets. We test the methods on two MMT architectures and find that they indeed make the model more reliant on the images.

2 Method

In this section, we elaborate two methods to conceal important information in the source textual inputs that can be recovered with the aid of visual inputs.

Back-Translation. Rather than trying to create reference translations that make use of visual signals for disambiguation, we treat original image captions as the target side and automatically produce ambiguous source sentences. While such back-translations are generally used for data augmentation (Sennrich et al., 2016), we rely fully on this data for training and testing. We focus on gender ambiguity, which can be easily created by translating from a language with natural gender (English) into a gender-neutral language (Turkish). In Turkish, there is no distinction between gender pronouns (*e.g.* "he" and "she" are both translated into "o"). We use a commercial translation system (Google Translate) to translate the image description in English to Turkish. The task is then to translate from Turkish back into English. An example is shown in Fig. 1.

Word Dropout. Inspired by Caglayan et al. (2019), we degrade the textual inputs to eliminate crucial information. We use a simplified approach that requires no manual annotation, randomly replacing tokens in the source sentence with a special UNK token, subject to a dropout probability p (Bowman et al., 2016).

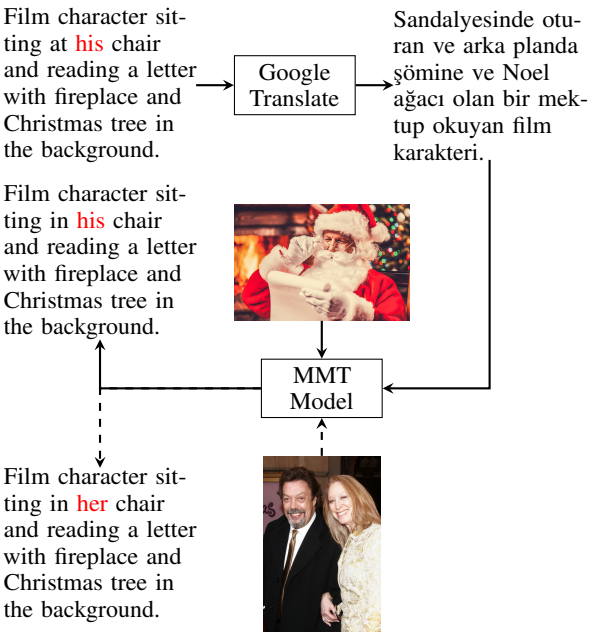

Figure 1: An example for back-translation. The image caption is translated into Turkish using a text-only translation system. Then a MMT model is trained to translate it back into English. When an incongruent image is fed into the model, the gender pronoun "his" is mistranslated.

3 Experimental Setup

3.1 Data Collection

As our starting point, we use Conceptual Captions (Sharma et al., 2018), which contains 3.3M images with captions. The captions in the dataset have already been processed to replace named entities with hypernyms such as 'person' or profession names such as 'actor'. In order to create a gender-ambiguous dataset we further filter out any sentences containing nouns with information about the gender of the entity (*e.g.* woman/man, lady/gentleman, king/queen, etc.) and also remove sentences with professions which are only used in a single gender-specific context (*e.g.* 'football player', which is always used with the male pronoun in the dataset). We then automatically translate the captions of the resulting dataset into Turkish and use this pseudo-parallel data for training our Turkish-English MMT models. For validation and testing we randomly sample 1000 sentences and use the remaining for training. We refer to this processed dataset as **Ambiguous Captions** (**AmbigCaps**).

For comparison, we also create a Turkish→English version of Multi30k by back-translating the English side. Tab. 1 summarizes the

characteristics of the two corpora.

Dataset	# Sen	# Words (EN)	# Gen. PROs
Multi30k	31,014	369,048	4,181
AmbigCaps	91,601	1,253,400	109,440

Table 1: Statistical properties (numbers of sentences, words, and gender pronouns in English) of the Multi30k and Ambiguous Captions datasets used in our experiments.

3.2 Models

In our experiments, we consider one NMT model and two MMT models. We follow Wu et al. (2021)'s model and configuration to isolate the cause for the negative results they obtained. We decide not to use the retrieval-based system because it samples images that are not described by the text. We also implement another simple model to demonstrate the applicability of our approaches.

Transformer. For text-only baseline, we use a variant of the Transformer that has 4 encoder layers, 4 decoder layers, and 4 attention heads in each layer. The dimensions of input/output layers and inner feed-forward layers are also reduced to 128 and 256 respectively. This configuration has been shown to be effective on Multi30K dataset (Wu et al., 2021). The MMT models below follow the same configuration.

Visual Features. Image features are extracted with the code snippet provided by Elliott et al. (2017),[2] which uses a ResNet-50 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009) as image encoder. The 'res4_relu' features $\in \mathbb{R}^{1024\times14\times14}$ and average pooled features $\in \mathbb{R}^{2048}$ are extracted.

Gated Fusion. Gated fusion model (Wu et al., 2021) learns a gate vector $\boldsymbol{\lambda}$, and combines textual representation and image representations as follows:

$$\mathbf{H} = \mathbf{H}_{\text{text}} + \boldsymbol{\lambda} \odot \mathbf{H}_{\text{avg}}, \tag{1}$$

where $\mathbf{H}_{\text{text}}$ is the output of the Transformer encoder, $\mathbf{H}_{\text{avg}}$ is the average pooled visual features after projection and broadcasting, and $\odot$ denotes the Hadamard product. $\mathbf{H}$ is then fed into the Transformer decoder as in NMT.

[2] https://github.com/multi30k/dataset/blob/master/scripts/feature-extractor

Concatenation. We implement a different approach to combine textual and visual features. The flattened and projected 'res4_relu' features $\mathbf{H}_{\text{res4\_relu}}$ are directly concatenated with the Transformer encoder representations $\mathbf{H}_{\text{text}}$ as follows:

$$\mathbf{H} = [\mathbf{H}_{\text{text}}; \mathbf{H}_{\text{res4\_relu}}]\,. \tag{2}$$

This preserves more fine-grained features in the original image and avoids confounding the two modalities.

3.3 Implementation Details

We follow (Wu et al., 2021) and use Adam (Kingma and Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. Maximum number of tokens in a mini-batch is 4096. Learning rate warms up from $1e-7$ to 0.005 in 2000 steps, then decays based on the the inverse square root of the update number. A dropout (Srivastava et al., 2014) of 0.3 and label-smoothing of 0.1 are applied. The models are trained with early-stopping (patience=10) and the last ten checkpoints are averaged for inference. We use beam search with beam size 5. We use the toolkit FAIRSEQ (Ott et al., 2019) for our implementation.

3.4 Metrics

BLEU. We compute the cumulative 4-gram BLEU scores (Papineni et al., 2002) to evaluate the overall quality of translation.

Gender Accuracy. Since we are most concerned with the gender ambiguity in the texts, we introduce gender accuracy as an additional metric. We first extract gender pronouns from the sentence. If the sentence contains at least one of the male pronouns ['he', 'him', 'his', 'himself'], it is classified as 'male'; if it contains at least one of the female pronouns ['she', 'her', 'hers', 'herself'], it is classified as 'female'; if it contains both male and female pronouns or neither, it is classified as 'undetermined'. We only consider the first two categories,[3] and compute gender accuracy by comparing the results of references and hypotheses.

Image Awareness. To examine models' reliance on the visual modality, we calculate the performance degradation when randomly sampled images are fed. This is also termed as image awareness (Elliott, 2018).

[3] See § 6.

4 Results

The results of our experiments are shown in Tab. 2.

4.1 Multi30K EN→DE

Test2016 We found our MMT models provide little to no improvement over the text-only Transformer. Moreover, the impact of feeding MMT systems with incongruent images is negligible. Our observations conform with previous work (Lala et al., 2018; Barrault et al., 2018; Wu et al., 2021), namely that visual signals are not utilized.

Multisense We also evaluate models trained on Multi30K on the Multisense test set (Gella et al., 2019). Similarly, no substantial difference is observed whether congruent or incongruent images are used. This suggests that it is not just a matter of the Test2016 test set containing too little textual ambiguity, but that the model has not learned to incorporate the visual information necessary for Multisense.[4]

4.2 Multi30K TR→EN

Our experiments on the TR→EN version of Multi30K that we created do not show any substantial improvements in image awareness, which we attribute to the relative sparsity of gender ambiguity at training and test time (see Tab. 1).

4.3 Ambiguous Captions

Training the same multimodal models on the Ambiguous Captions dataset results in substantial improvements in terms of both BLEU scores and gender accuracy compared to our text-only baseline. This suggests that the high level of textual ambiguity in this dataset encourages MMT models to exploit visual information. We further test this hypothesis by repeating the experiment when images are shuffled, and observe that their performance substantially deteriorates, especially their ability to infer the correct gender pronouns. For instance, the gated fusion model has an impressive gender accuracy of 80.9% compared to 73.9% of the text-only Transformer, while it drops to 64.4% when incongruent images are used.

We find that both the gated fusion and concatenation model behave similarly, indicating that the choice of dataset has a bigger effect on the success of multimodal modeling than the specific architecture.

[4]We also note that some senses in Multisense are rare or unseen in Multi30k.

4.4 Effect of Word Dropout

We found word dropout tends to increase image awareness for the concatenation model. This is most evident for Multi30K (TR→EN), where image awareness increases by $\approx 300\%$. For the gated fusion model, although word dropout leads to more differences in translations between congruent and incongruent image-text alignments (e.g. on Multi30K (TR→EN), 20 differences without dropout, 192 with dropout), it is not well reflected by the image awareness metric. The reason remains to be further inspected.

Despite having the desired effect of increasing image awareness on the concatenation model, we observe some deterioration of BLEU and gender accuracy compared to the model trained without word dropout; still, we hope that our results serve as a proof-of-concept to motivate future research on regularization schemes that aim to (re)balance visual and textual signal. We note the success of work done in parallel to ours that applied word dropout to increase context usage in context-aware machine translation (Fernandes et al., 2021).

5 Conclusion

Our experiments explain recent failures in MMT, and show that the models we examine successfully learn to rely more on images when textual ambiguity is high (as in our back-translated Turkish–English dataset) or when textual information is dropped out. Our results suggest that simple MMT models have some capacity to integrate visual and textual information, but their effectiveness is hidden when training on datasets where the visual signal provides little information. In the long term, we hope to identify real-world applications where multimodal context naturally provides a strong disambiguation signal. For the near future, we release our dataset and encourage researchers to utilize it to validate future research on multimodal translation models. For example, we are interested under which conditions multimodal models learn to exploit visual signal: does the absolute frequency of examples with textual ambiguity matter more, or their proportion?

6 Broader Impact Statement

Our dataset inherits biases from the Conceptual Captions dataset. We cannot rule out gender bias in the dataset similar to the one described by Zhao

Model	Multi30K (EN→DE)		Multi30K (TR→EN)		Ambiguous Captions	
	Test2016 BLEU	Multisense BLEU	BLEU	Gender Accuracy	BLEU	Gender Accuracy
Transformer	40.53	26.65	51.64	67.0%	35.71	73.9%
Gated Fusion	41.22 (↑ 0.01)	27.09 (↓ 0.04)	51.76 (↑ 0.04)	72.2% (↓ 0.5%)	36.68 (↓ 1.71)	80.9% (↓ 16.5%)
+ Word Dropout	40.65 (↓ 0.19)	26.09 (↑ 0.15)	51.07 (↑ 0.06)	66.1% (↑ 0.5%)	35.35 (↓ 1.28)	79.3% (↓ 16.1%)
Concatenation	39.86 (↑ 0.02)	25.71 (↑ 0.25)	51.34 (↓ 0.25)	72.2% (↑ 1.4%)	37.39 (↓ 2.08)	79.4% (↓ 18.1%)
+ Word Dropout	40.07 (↓ 0.50)	25.72 (↓ 0.07)	50.81 (↓ 0.90)	68.7% (↓ 3.5%)	35.55 (↓ 2.10)	79.0% (↓ 18.2%)

Table 2: Models' performance on various datasets. In the parenthesis is the drop when incongruent images are used (i.e. image awareness). We take the average of 5 runs, each with a different random seed. ↑ indicates the performance improves after the images are shuffled; ↓ otherwise.

et al. (2017), with males and females showing different distributions, and we only studied a subset of captions with unambiguously male or female pronouns. Despite potential issues with our dataset (which we consider unsuitable for use in production because of aggressive filtering), we believe our work on improving MMT has a positive effect on gender fairness, since multimodal systems with audiovisual clues have the potential to reduce gender bias compared to systems that only rely on textual co-occurrence frequencies.

Acknowledgments

We would like to thank the anonymous reviewers and meta-reviewer for their comments. This project has received support from the Swiss National Science Foundation (MUTAMUR; no. 176727).

References

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing

context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online. Association for Computational Linguistics.

Spandana Gella, Desmond Elliott, and Frank Keller. 2019. Cross-lingual visual verb sense disambiguation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1998–2004, Minneapolis, Minnesota. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Jindřich Helcl, Jindřich Libovický, and Dušan Variš. 2018. CUNI system for the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 616–623, Belgium, Brussels. Association for Computational Linguistics.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538, Florence, Italy. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.

Chiraag Lala, Pranava Swaroop Madhyastha, Carolina Scarton, and Lucia Specia. 2018. Sheffield submissions for WMT18 multimodal translation shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 624–631, Belgium, Brussels. Association for Computational Linguistics.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1320–1329, New York, NY, USA. Association for Computing Machinery.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Vikas Raunak, Sang Keun Choe, Quanyang Lu, Yi Xu, and Florian Metze. 2019. On leveraging the visual modality for neural machine translation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 147–151, Tokyo, Japan. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Josiah Wang, Pranava Madhyastha, Josiel Figueiredo, Chiraag Lala, and Lucia Specia. 2021. Multisubs: A large-scale multimodal and multilingual dataset. *CoRR*, abs/2103.01910. Version 2.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6153–6166, Online. Association for Computational Linguistics.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In *Proceedings*

of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3025–3035, Online. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.